DOCUMENT RESUME

ED 236 238 TM 830 765

AUTHOR Shannon, Gregory A.
TITLE Objective-Referenced-Test Rescore Decisions and Item Statistics: A Matter of Congruence.
SPONS AGENCY Educational Testing Service, Princeton, N.J. Center for Occupational and Professional Assessment.
PUB DATE Apr 83
NOTE 25p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, Montreal, Quebec, April 11-15, 1983).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Adults; *Criterion Referenced Tests; *Item Analysis; *Occupational Tests; Scoring; *Testing Problems; Testing Programs; *Test Items
IDENTIFIERS Biserial Correlation; *Flawed Items; Item Discrimination (Tests); *Test Rescoring

ABSTRACT

Rescoring of Center for Occupational and Professional Assessment objective-referenced tests is decided largely by content experts selected by client organizations. A few of the test items, statistically flagged for review, are not rescored. Some of this incongruence could be due to the use of the biserial correlation (r-biserial) as an item-discrimination index. Item-flagging outcomes generated using the present item-discrimination index, r-biserial, were compared with those generated using phi/phi max and phi. The analysis sample consisted of responses of 10,702 candidates to 1,794 items taken from 23 different test forms. Major findings suggest that the incongruent items, initially flagged using r-biserial, tended also to be flagged when either phi/phi max or phi were employed. The item-rescore decisions made by the content experts were unsupported by the results, suggesting that their actions should be more closely monitored. Some implications for use of the three item-discrimination indices are also discussed. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED236238

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☒ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

OBJECTIVE-REFERENCED-TEST RESCORE DECISIONS AND ITEM STATISTICS: A MATTER OF CONGRUENCE

Gregory A. Shannon

Educational Testing Service

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

G.A. Shannon

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

A paper presented at the Annual Meeting of the
American Educational Research Association
Montreal, Canada, April 1983.

TM830765

This project was partially funded by the Center for Occupational and Professional Assessment of Educational Testing Service.

PRINTED IN U.S.A.

GENERAL FOOTNOTE: PAGE 1

This study was partially funded by the Center for Occupational and Professional Assessment, Educational Testing Service. A special note of appreciation is extended to Dr. Gerald J. Melican and Ms. Barbara A. Carroll, both of Educational Testing Service, for their valuable assistance with data collection and analysis.

OBJECTIVE-REFERENCED-TEST RESCORE DECISIONS AND ITEM STATISTICS: A MATTER OF CONGRUENCE

Gregory A. Shannon

Educational Testing Service

Rescoring of Center for Occupational and Professional Assessment objective-referenced tests is decided largely by content experts selected by client organizations. A few of the test items, statistically flagged for review, are not rescored. Some of this incongruence could be due to the use of the biserial correlation (r-biserial) as an item-discrimination index. Item-flagging outcomes generated using the present item-discrimination index, r-biserial, were compared with those generated using phi/phi max and phi. The analysis sample consisted of responses of 10,702 candidates to 1,794 items taken from 23 different test forms. Major findings suggest that the incongruent items, initially flagged using r-biserial, tended also to be flagged when either phi/phi max or phi were employed. The item-rescore decisions made by the content experts were unsupported by the results, suggesting that their actions should be more closely monitored. Some implications for use of the three item-discrimination indices are also discussed.

Author Abstract

TM 830 765

OBJECTIVE-REFERENCED-TEST RESCORE DECISIONS AND ITEM STATISTICS: A MATTER OF CONGRUENCE

Rescoring of objective-referenced tests administered by the Center for Occupational and Professional Assessment (COPA) division of Educational Testing Service (ETS) is decided largely by committees of examiners subject-matter experts selected by the client organization. Indeed, the role of the content experts in the development and revision of objective-referenced tests is paramount. Still, one must consider the dilemma of psychometricians who select (flag) possible problem items via statistical analyses only to have content experts indicate that the flagged items do not require rescoring. Rescoring refers to the process of correcting item answer keys following the initial scoring phase.

For most COPA exams, fewer than 3% of the items are rescored. Of these, about a third would be considered incongruent. Incongruent items are items that are statistically flagged, but not rescored. Some of the incongruence could be due to the use of the biserial correlation (r-biserial) as an item-discrimination index. Although r-biserial has been generally considered by ETS to be appropriate for norm-referenced and objectives-referenced tests, this index is more appropriate for measuring item discrimination along a total score range than at a specified passing score.

COPA tests are principally used to assign scores of PASS or FAIL to candidates for licensure and certification in more than 20 vocational fields, such as podiatry, electronic data processing, and real estate. Would flagging outcomes have been different if items had been flagged on the basis of indices measuring item discrimination at specified passing scores?

Numerous measures have been offered as item-discrimination indices for objectives-referenced tests. Indices that seem to be in wide use include: upper-lower indices, such as the B and D indices (Brennan, 1972) and correlational indices such as phi (Aiken, 1979 and Hsu, 1971) and phi/phi max (Carroll, 1961 and Cureton, 1959), among others. Since r-biserial can be considered to be a correlational index, it seemed reasonable to compare it with objectives-referenced item-discrimination indices considered also to be correlational. The indices selected were phi, because of its wide use, and phi/phi max, because it had been offered as a means of compensating for the conditions that restrict the size of phi (Cureton, 1959). In the objectives-referenced test sense, phi can attain the maximum values of +1 or -1 only when the number of people passing the item equals the number of people passing the test. This problem has been discussed by Cureton, 1959 and Glass and Stanley, 1970. It is this author's contention that phi and phi/phi max appear to be potentially useful correlational indices for identifying possible problem test items for objectives-referenced tests.

In the present study, item-flagging outcomes generated using r-biserial (considered to be more appropriate for norm-referenced tests) will be compared with those generated using phi/phi max and phi. And even more important, would the examination committees' decisions not to rescore the items flagged using r-biserial (incongruent items) be supported by the flagging outcomes based upon phi/phi max and phi?

METHOD

At COPA, test items undergo several stages of content and statistical review, both by internal statistical and test development staff and by external content experts (e.g., committees of examiners). A preliminary item analysis (PIA) is performed for all items on each new test form, shortly after answer sheets have been scanned. The ETS generalized item analysis program, IANA-10, is used for this purpose. By using PIA output, the statistical staff flags items on the basis of statistical information such as r-biserials, item difficulty indices (deltas and percentage of candidates who answered each item correctly (p-values)), item-choice distributions, and mean ability scores (estimated from test scores) of candidates who chose each response. Biserial correlations are calculated by first transforming each raw total test score distribution to a normal distribution, and then setting the mean and standard deviation to 13 and 4, respectively. The resulting scores serve as criterion scores for calculating r-biserials, to satisfy the assumption that criterion scores should be normally distributed. For COPA programs in the study, in the case of

item discrimination, items with r-biserials less than .10 are flagged for review. A list of flagged items is then reviewed by committees of examiners, who, as content experts, decide if any referred items are to be rescored. The rescore decisions of these content experts usually stand, unless strong objections are made by COPA statistical, test development, or management staff. Rescoring is followed by final item analysis, score equating and scaling, and score reporting, as specified for each program.

The present study focuses upon the PIA phase and the period that follows, when examination committees decide which (if any) of the statistically flagged items are to be rescored. For the purpose of the present study, 23 testing administrations were sampled from those that occurred between April 1981 and April 1982. In selecting the sample, an effort was made to select administrations for which one or more test items were incongruent (had been statistically flagged using r-biserial, but had not been identified for rescoring by the examination committees).

The six testing programs selected for the present study administer tests to licensure and certification candidates at testing centers located throughout the contiguous United States (with Program B also testing a small number of candidates in foreign centers). In the case of all programs, tests are scored as "rights only", as opposed to being corrected for guessing. All test items are multiple-choice, ranging from four choices (Programs B, C, D, and F) to five choices (Programs A and E). Programs A through E are certification programs, while Program F is a licensure program.

R-biserials were taken from the PIA output. In order to obtain statistics that more directly address objectives-referenced tests, two additional computer programs were employed. An existing Speakeasy scaling program was used to estimate the raw cut scores, required to calculate phi/phi and phi. The method of scaling was to set the means and standard deviations of the new forms to those of the old forms (forms selected to serve as old forms for score equating). New form statistics were taken from PIA output, while old form statistics were taken from final item analysis output generated when the old forms were first administered. The raw scores of the new forms that scaled to the cut scores of the old form served as the estimated raw cut scores.

The cut scores that had been estimated using the first program served as input for a Fortran program written for the present study. This second program read the scanning tapes that had been read by the PIA program. The program calculated the item-discrimination indices (phi/phi max and phi), p-values, raw score means and standard deviations (for quality control purposes), and percentage of candidates passing the test.

RESULTS

Data were collected for testing administrations of 23 different test forms, administered by the six testing programs. Results are summarized in Table 1. The testing administrations were assigned three-value codes, for purposes of anonymity. The first value of the code is the testing program code (e.g., Program A), the second value

is the test code (e.g., Test 1), and the third value is the test form code (Form 1).

The sample consisted of 1,838 test items administered to 10,702 candidates tested between April 1981 and April 1982. The lengths of the administered tests ranged from 30 to 400 items; and the number of candidates tested during a single administration ranged from 76 to 1,182. Percentages of candidates who passed the tests are based upon cut scores estimated for preliminary item analysis purposes; they deviate only slightly from the passing rates based upon reported cut scores (subsequently identified following item rescoring, final item analysis, and statistical equating (for most programs)). The percent passing ranged from 21% to 95% for individual testing administrations.

Except for Program C (which samples the total candidate group for PIA), only the scores of first-time test takers (with repeaters excluded) had been included in preliminary item analyses. Another point of clarification concerns Program F, where candidates who had taken one given test are identical to those who had taken another test. The preliminary item analysis program had been programmed to sample only those candidates who had taken both Tests 1 and 2, and to sample only those candidates who had taken both Tests 3 and 4. This is because Tests 1 and 2 are interpreted as adjacent measures of one knowlege area, while Tests 3 and 4 are interpreted as adjacent measures of another knowlege area.

Of the 1,838 items processed, 44 were eliminated on the basis of being out of range. That is, 44 items had been correctly answered by more than 95% or less than 5% of the candidates tested, to avoid the

problem of interpreting r-biserials based upon minimum item variance, due to extreme values of item difficulty. This rule is consistent with that defined for the PIA program. The remaining 1,794 items served as the analysis sample.

Means, standard deviations, and ranges of the indices of item discrimination (r-biserial, phi/phi max, and phi) and difficulty (p-values) are presented in Table 2. The r-biserial and phi/phi max means are close, although more variation is observed for phi/phi max. The mean of phi is much lower than those observed for the other discrimination indices, probably due largely to tendency of phi to be restricted in size, as the number of correct item responses departs from the number of candidates that passed the test. Given the wide range of passing rates and item difficulty, it would not be uncommon to expect many instances of restricted values of phi. The mean p-value is .646, suggesting that on the average, the items were answered correctly by most of the candidates in the sample.

Presented in Table 3 are crosstabulations of congruency states (defined in terms of r-biserial), with item-flagging outcomes identified for phi/phi max and phi. Congruency states are examination committee actions taken in reponse to r-biserial item-flagging outcomes. Of the 158 items flagged using r-biserial indices, 134 were incongruent (flagged using r-biserials but not rescored). Items defined as congruent are the 1,636 items that were not flagged and not rescored, and the 24 items that were flagged and rescored. Of the 134 incongruent items, 115 (86%) were also flagged using phi/phi max and 130 (97%) were also flagged using phi. These findings suggest that the incongruent items, initially flagged using r-biserial, tended also

to be flagged when either phi/phi max or phi were employed.

Of the 1,636 congruent items not flagged using r-biserial and not rescored, 231 would have been flagged using phi, but only 92 would have been flagged using phi/phi max. This is probably due largely to restrictions in the size of phi.

Crosstabulations of item-flagging outcomes identified for the three discrimination indices are presented in Table 4. The largest number of items were flagged using phi (383 items), followed by phi/phi max (227 items), and r-biserial (158 items). By considering the percentage of all 1,794 items for which the flagging outcomes were in agreement, one could estimate the extent that a given pair of coefficients yielded the same flagging decisions. For example, 135 items were flagged using both r-biserial and phi/phi max, while 1544 items were not flagged using either of these indices. By summing these frequencies (135+1544=1679), dividing by 1794 (1679/1794=.936), and multiplying by 100% (.936*100%=94%), one could calculate the percentage of agreement between flagging outcomes of a given pair of discrimination indices. The percentage of agreement of flagging outcomes is 94% between r-biserial and phi/phi max, 91% between phi/phi max and phi, and 87% between phi and r-biserial. This suggests that the three discrimination indices tended to yield similar flagging outcomes, with agreement being highest between phi/phi max and r-biserial.

Of the 158 items flagged using r-biserial, 135 (60%) were flagged using phi/phi max and 152 (96%) were flagged using phi. These results suggest that the tendency of phi to become restricted in magnitude due

to extreme marginals may also result in a large, perhaps excessive, number of items being flagged when phi is employed as an item-discrimination index.

The Product-Moment correlations among the three discrimination indices, presented in Table 5, tend to be from moderate to high, ranging from .671 (phi with phi/phi max) to .834 (phi with r-biserial). Correlations between indices of item difficulty and the three discrimination indices range from .178 (with phi), to .238 (with phi/phi max), to .320 (with r-biserial). Effective item-discrimination indices should not be correlated with item difficulty.

For further comparisons of item-discrimination indices, the 23 test administrations described in Table 1 were partitioned into three categories, according to distance of scaled cut scores from scaled total score mean. Level 1 (hardest) includes test administrations for which the cut score was more than 1/3 of a standard deviation above the test mean; Level 2 (middle) includes test administrations for which the cut score was within 1/3 of a standard deviation of the test mean; and Level 3 (easiest) includes test administrations for which the cut score was more than 1/3 of a standard deviation below the test mean. Assignments of test administrations to levels, in addition to scaled cut scores in standard deviation units, appear in Table 6. The number of administrations assigned to levels is 9 to Level 1, 6 to Level 2, and 8 to Level 3. It should be noted that Level 3 consists primarily of Program D test administrations. More than half of the scaled cut scores, 13 or 57%, fall below the scaled total score means, characteristic of COPA testing programs. The spread of the deviation

scores is not wide. With the exception of F31 and F41, all administrations had been assigned cut scores that fell within one standard deviation of the scaled total score means.

Indices of item-discrimination and difficulty are summarized statistically in Table 7, by level of leniency. The largest concentration of items is in Level 2, 778 or 43%. The highest mean (.458) is observed for phi/phi max at Level 1, possibly because the discrepancy between the numbers of candidates passing the item and the numbers of candidates passing the test may have been prominent at this level, as suggested by the discrepancy between the mean p-value and the percent passing.

Mean item difficulty appears to be reasonably close across levels, ranging from .621 (Level 2) to .676 (Level 1). The highest mean p-value was observed at Level 1, suggesting that there may have been a tendency for the Test Assemblers to select easier items to compensate somewhat for the low passing rates associated with this level.

Percentages of agreement among flagging outcomes, discussed previously in reference to Table 4, are presented in Table 8 by level of leniency. The reasonably high agreement observed for the total group of items (N=1794), which ranges from 87% to 94%, deviates only slightly from the percentages reported at each level. At none of the levels do the percentages of agreement fall below 85%.

DISCUSSION

This study was conducted to investigate the incongruent circumstance of items being flagged because of low biserials that are not rescored by committees of examiners. It seemed possible that the actions of these content experts might have been more consistent with flagging outcomes had more objectives-referenced item-discrimination indices been used. Such support was not evident from the results. Although consistency was high between the r-biserial flagging results and those of phi/phi max and phi (for both incongruent and congruent items), the reader should not conclude that r-biserials would be appropriate as item-discrimination indices for tests used to assign people into categories of PASS or FAIL. Biserials can be useful for discriminating between item scores and criterion scores, along the range of criterion scores, rather than at the score that separates passing and failing test takers. The high agreement observed in the present study was probably influenced largely by the fact that the cut scores defined for the tests tend to be located close to the total test means, where the similarity between r-biserial and objectives-referenced indices such as phi/phi max in particular would probably be maximum. This is also suggested by the comparisons of mean indices at three cut score levels.

The above findings suggest that the actions of the committees of examiners should be more closely monitored by both Test Development and Statistical staff. For the present sample, r-biserial appeared to be useful as an item-discrimination index for flagging items on objectives-referenced tests. However, for conceptual reasons, this author does not consider r-biserial to be a logical choice for this

purpose, unless cut scores are equal to test score means. In practice, distances between cut scores and test score means have been observed to vary widely, frequently more widely than the distances reported for the present sample. In the interest of prudence, one should select an index that can be expected to discriminate at cut scores, regardless of their location in relation to mean test scores. For this reason, it is recommended that r-biserial be replaced with an index that is more appropriate for objectives-referenced tests. (A special study should be conducted to select the replacement(s).) Examination committees should also be provided with computer printouts that detail for each item the differential responses of passing and failing candidates. Committees might then be in a better position to weigh statistical information in making rescoring decisions. In addition, such explicit information could also serve as a starting point for instructing examination committees about assembling, evaluating, and rescoring COPA tests.

Another point of interest concerns the use of phi for flagging test items. The problem of phi being restricted in size when the number passing the test and the number passing the item are discrepant also seemed evident when phi was used to flag test items. For example, although the overall agreement was moderately high between the phi and r-biserial flagging outcomes, more than twice as many items were flagged using phi as using r-biserial. A psychometrician interested in replacing r-biserial with phi might be inclined to simply raise the standard for flagging items (e.g., from .10 to .15). Such a practice is not recommended since the degree to which phi would be restricted in size would depend upon the discrepancies of the

marginals, which can be expected to vary depending upon the number passing the test and the item. For example, it would not be reasonable to assume that the degree of restrictions upon the size of phi would be the same for items with phi values of .10 as they would be for items with phi values of .15. In view of the potential for misinterpretation and misuse by practitioners, for flagging purposes at least, it would be more advisable to use phi/phi max, or an upper-lower index such as the B-index (Brennan, 1972).

A desirable property of the three discrimination indices examined is that none of them appeared to be correlated with item difficulty, although some readers might consider the correlation between the r-biserials and the p-values (r=.32) to be somewhat less than comforting. If an item discrimation index is correlated with item difficulty, interpretation of the index could become a problem.

REFERENCES

Aiken, L. R. Relationships between the item difficulty and discrimination indexes. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1979, 39, 821-824.

Brennan, R. L. A generalized upper-lower item discrimination index. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1972, 32, 289-303.

Carroll, J. B. The nature of data, or how to choose a correlation coefficient. PSYCHOMETRIKA, 1961, 26, 347-372.

Cureton, E. E. Note on phi/phi max. PSYCHOMETRIKA, 1959, 24, 89-91.

Glass, G. V and Stanley, J. C. STATISTICAL METHODS IN EDUCATION AND PSYCHOLOGY. Englewood Cliffs: Prentice-Hall, Inc., 1970, 158-162.

Hsu, T. Empirical data on criterion-referenced tests. Paper presented at the Annual Meeting of the American Educational Research Association, New York, New York, February, 1971.

Table 1

Summary Statistics Reported for the COPA Test Administrations in the Analysis

Adm. Code[a]	Adm. Date	No. Tested	No. of Items	Raw Score Mean	Raw Score SD	Raw Score Cut	Scaled Score Mean	Scaled Score SD	Scaled Score Cut	Percent Passing
A11	06/81	1030	180	129.8	14.5	116	126.7	13.0	114	85
B11	04/82	629	400	248.0	34.3	241	261.0	35.2	254	66
C21	06/81	948	70	44.8	8.2	42	57.0	5.8	55	67
C31	06/81	1067	50	27.6	5.9	27	55.5	5.1	55	56
C41	06/81	959	40	23.7	4.4	22	57.4	6.2	55	69
C12	01/82	1182	80	48.8	8.6	46	57.1	7.1	55	67
C22	01/82	1101	70	43.0	7.2	40	57.0	5.8	55	71
C32	01/82	953	50	30.4	5.8	30	55.4	5.3	55	60
C42	01/82	810	60	34.8	6.2	36	54.0	5.2	55	49
D11	04/81	157	50	36.2	6.6	39	52.9	5.1	55	38
D21	04/81	158	50	35.2	5.2	39	52.1	4.0	55	31
D31	04/81	176	50	34.6	5.7	39	51.6	4.4	55	23
D41	04/81	205	50	32.9	6.2	39	50.3	4.7	55	21
D12	10/81	149	50	35.7	6.4	39	52.4	4.9	55	37
D22	10/81	134	50	34.8	5.5	39	51.8	4.3	55	29
D32	10/81	153	50	34.7	5.8	39	51.7	4.4	55	28
D42	10/81	191	50	33.7	6.0	39	50.9	4.6	55	24
E11	11/81	160	150	97.0	15.3	102	66.9	8.8	70	41
E21	11/81	112	150	103.4	10.9	100	71.8	6.0	70	70
F11	02/82	138	34	21.0	4.4	18	76.9	8.8	70	79
F21	02/82	138	39	24.8	4.9	20	74.8	5.0	70	86
F31	02/82	76	30	20.2	2.6	17	74.7	4.1	70	95
F41	02/82	76	35	22.7	3.2	19	75.6	4.9	70	92
A11		10702	1838							62

[a] The administration code is defined as follows:
first value = testing program code, e.g., Program A;
second value = test code, e.g., Test 1; and
third value = test form code, e.g., Form 1.

Table 2

Summary Statistics for Indices of Item Discrimination And Difficulty

Index	Mean	SD	Min.[a]	Max.[b]	No. Tested	No. of Items	Percent Passing
R-Bis	.311	.154	-.250	.840	10702	1794	62
P/PMax	.321	.236	-1.000	1.000	10702	1794	62
Phi	.175	.098	-.157	.587	10702	1794	62
P-Value	.646	.197	.051	.949	10702	1794	62

[a] Minimum observed value of index.

[b] Maximum observed value of index.

Table 3

Crosstabulations of Congruency States (Defined in Terms of R-Biserial) with Item-flagging Outcomes Identified for The Other Two Discrimination Indices

		Congruent: Not Rescored	Congruent: Rescored	Incongruent: Not Rescored	Totals
Phi/ Phi Max	Flagged	92 [a]	20	115	227
		40.5	8.8	50.7	12.7
		5.6	83.3	85.8	
	Not Flagged	1544	4	19	1567
		98.5	0.3	1.2	87.3
		94.4	16.7	14.2	
	Totals	1636	24	134	1794
		91.2	1.3	7.5	100.0

		Congruent: Not Rescored	Congruent: Rescored	Incongruent: Not Rescored	Totals
Phi	Flagged	231	22	130	383
		60.3	5.7	33.9	21.3
		14.1	91.7	97.0	
	Not Flagged	1405	2	4	1411
		99.6	0.1	0.3	78.7
		85.9	8.3	3.0	
	Totals	1636	24	134	1794
		91.2	1.3	7.5	100.0

NOTE. Congruency states are examination committee actions taken in response to R-BISERIAL item-flagging outcomes. The states are defined as follows:

Congruent: Not Rescored = 1,636 items not flagged (using r-biserial) and not rescored,

Congruent: Rescored = 24 items flagged (using r-biserial) and rescored, and

Incongruent: Not Rescored = 134 items flagged (using r-biserial) and not rescored.

[a] The values in each cell are arranged as follows:
row 1 = cell frequency,
row 2 = percentage of row total, and
row 3 = percentage of column total.

Table 4

Crosstabulations of Item-flagging Outcomes
Identified for The Three Discrimination Indices

		Phi/Phi Max Flagged	Phi/Phi Max Not Flagged	Totals
R-Bis	Flagged	135 [a]	23	158
		85.4	14.6	8.8
		59.5	1.5	
	Not Flagged	92	1544	1636
		5.6	94.4	91.2
		40.5	98.5	
	Totals	227	1567	1794
		12.7	87.3	100.0

		Phi Flagged	Phi Not Flagged	Totals
Phi/ Phi Max	Flagged	227	0	227
		100.0	0.0	12.7
		59.3	0.0	
	Not Flagged	156	1411	1567
		10.0	90.0	87.3
		40.7	100.0	
	Totals	383	1411	1794
		21.3	78.7	100.0

		R-Bis Flagged	R-Bis Not Flagged	Totals
Phi	Flagged	152	231	383
		39.7	60.3	21.3
		96.2	14.1	
	Not Flagged	6	1405	1411
		0.4	99.6	78.7
		3.8	85.9	
	Totals	158	1636	1794
		8.8	91.2	100.0

[a] The values in each cell are arranged as follows:
row 1 = cell frequency,
row 2 = percentage of row total, and
row 3 = percentage of column total.

Table 5

Correlations Among Indices of Item Discrimination And Difficulty

	R-Bis	P/PMax	Phi	P-Value
R-Bis	1.000			
P/PMax	.755	1.000		
Phi	.834	.671	1.000	
P-Value	.320	.238	.178	1.000

NOTE. All correlations are based on an analysis sample of 1,794 test items.

Table 6

Scaled Cut Scores Classified into Three Levels of Leniency

Adm. Code[a]	Scaled Score Mean	Scaled Score Cut	Scaled Cut in SD Units	Level of Leniency[b]
F31	74.7	70	-1.15	3
F41	75.6	70	-1.14	3
A11	126.7	114	- .97	3
F21	74.8	70	- .95	3
F11	76.9	70	- .78	3
C41	57.4	55	- .38	3
C22	57.0	55	- .35	3
C21	57.0	55	- .34	3
E21	71.8	70	- .30	2
C12	57.1	55	- .29	2
B11	261.0	254	- .20	2
C31	55.5	55	- .10	2
C32	54.4	55	- .09	2
C42	54.0	55	.20	2
E11	66.9	70	.35	1
D11	52.9	55	.42	1
D12	52.4	55	.52	1
D21	52.1	55	.74	1
D32	51.7	55	.75	1
D22	51.8	55	.75	1
D31	51.6	55	.78	1
D42	50.9	55	.90	1
D41	50.3	55	1.00	1

[a] The administration code is defined as follows:
first value = testing program code, e.g., Program A;
second value = test code, e.g., Test 1; and
third value = test form code, e.g., Form 1.

[b]
Level 3 (Easiest) = test administrations for which cut score was more than 1/3 SD below scaled total test mean.
Level 2 (Middle) = test administrations for which cut score was within 1/3 SD of scaled total test mean.
Level 1 (Hardest) = test administrations for which cut score was more than 1/3 SD above scaled total test mean.

Table 7

Summary Statistics for Indices of Item Discrimination
And Difficulty by Level of Leniency

Level of Leniency[a]	No. Tested	No. of Items	R-Bis Mean	R-Bis SD	P/PMax Mean	P/PMax SD	Phi Mean	Phi SD	P-Value Mean	P-Value SD	Percent Passing
3	4466	479	.321	.151	.291	.237	.177	.102	.652	.205	75
2	4753	778	.270	.145	.244	.167	.164	.095	.621	.199	60
1	1483	537	.362	.154	.458	.262	.188	.097	.676	.182	30
Total	10702	1794	.311	.154	.321	.236	.175	.098	.646	.197	62

[a]
Level 3 (Easiest) = test administrations for which cut score was more than 1/3 SD below scaled total test mean.

Level 2 (Middle) = test administrations for which cut score was within 1/3 SD of scaled total test mean.

Level 1 (Hardest) = test administrations for which cut score was more than 1/3 SD above scaled total test mean.

Table 8

Percentages of Agreement among Flagging Outcomes Based on Three Indices of Item Discrimination by Level of Leniency

Level of Leniency[a]	No. of Items	R-bis with Phi/Phi Max	R-Bis with Phi	Phi/Phi Max with Phi
3	476	92	91	85
2	778	93	93	88
1	537	95	90	86
Total	1794	94	91	87

[a] Level 3 (Easiest) = test administrations for which cut score was more than 1/3 SD below scaled total test mean.

Level 2 (Middle) = test administrations for which cut score was within 1/3 SD of scaled total test mean.

Level 1 (Hardest) = test administrations for which cut score was more than 1/3 SD above scaled total test mean.